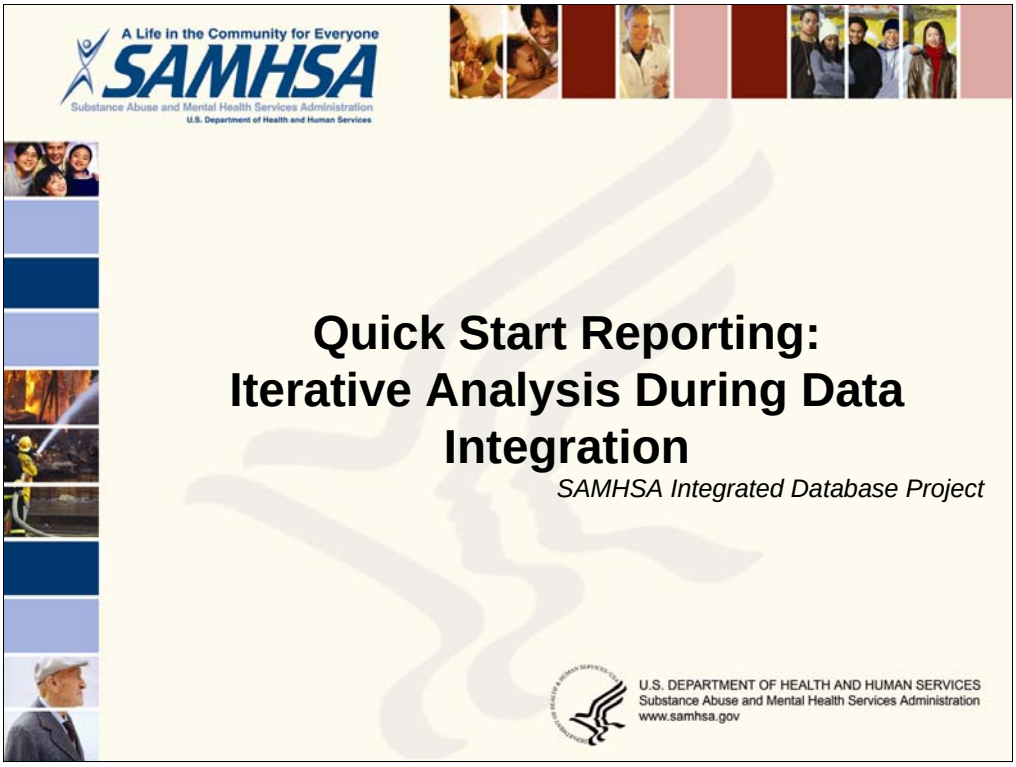
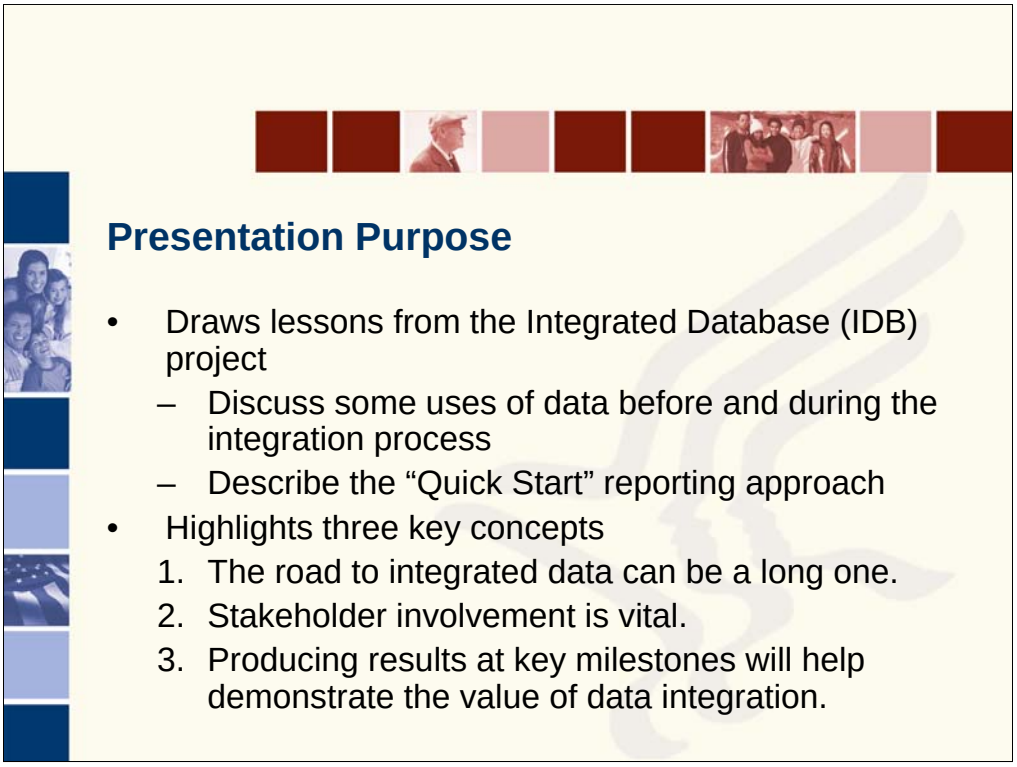


Hello, and welcome to “Quick Start: Reporting During Data Integration.” This presentation was developed for SAMHSA’s Integrated Database Project as part of a broader effort to share information on:

- Integrating mental health and substance abuse administrative data with other data sources, such as Medicaid data
- Using that integrated data to improve operations and inform policy decisions.

SAMHSA’s Integrated Database (or IDB) Project is an effort jointly funded by SAMHSA’s Center for Mental Health Services and SAMHSA’s Center for Substance Abuse Treatment. And, by the way, we refer to this work as the “IDB Project”. However, the official name of the contract is *Integrated Medicaid and State Agency Mental Health and Substance Abuse: Data Analysis and Technical Assistance*



This presentation draws lessons from Medstat’s IDB project. Notably, that rapidly creating useful analyses and reports is an important component of any data integration process. With this presentation, we hope to displace the notion that agencies must wait until fully integrated data is available before generating valuable information. Instead, we want to impress upon you that useful reports are both possible and actually necessary throughout the process of integrating data.

This presentation explains the important points, or milestones, of the data integration process, and reveals the type of achievable and important work appropriate at each of these stages. Remember that the real goal is not just creating an integrated database. The data is a means for informing policy and operations, as well as improving services. The reports and analyses we describe are helpful examples that can do exactly these things. The information also serves another important purpose: reminding stakeholders that useful and important work is underway.

This iterative approach will facilitate data integration by allowing these constituents to actively participate in the review and analysis of interim data. We refer to this approach as “Quick Start” Reporting.

The points we hope you will take away:

- Data integration takes a lot of time, even in the most favorable situations.
- Stakeholder participation is vital. Producing results during the integration process is a useful way of maintaining stakeholder involvement.
- Creating results throughout the process demonstrates value

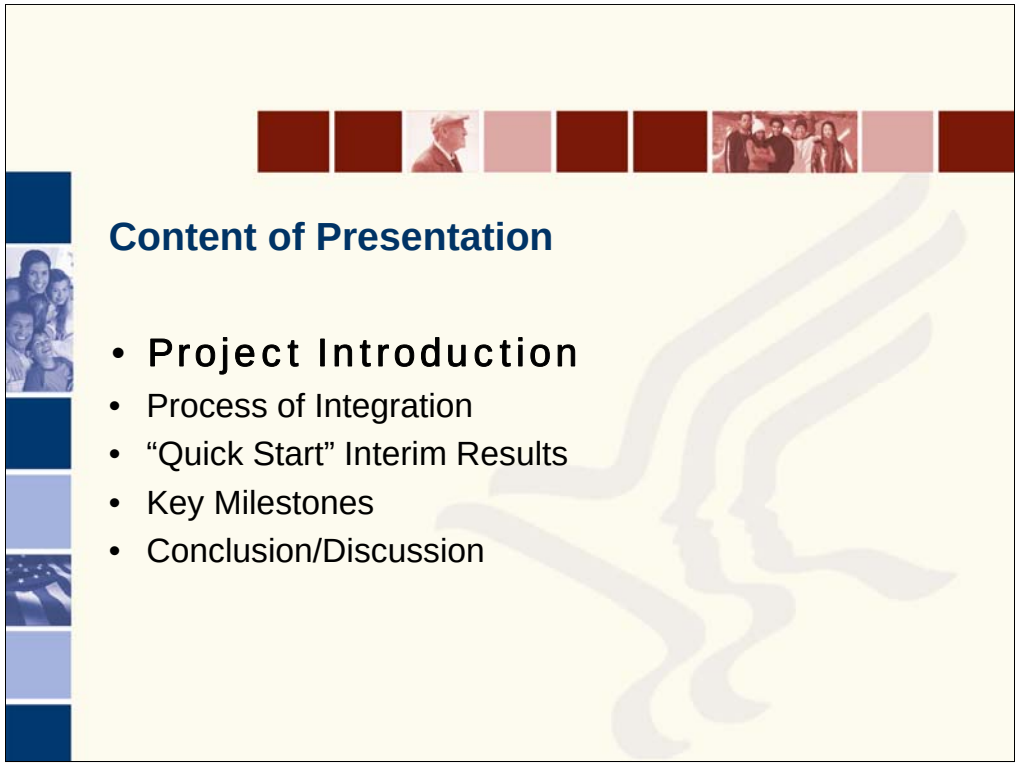
This is a straightforward presentation designed to.

- Lay out a simple process
- Mark milestones along that process, and
- Describe the types of work that can be accomplished at each milestone.

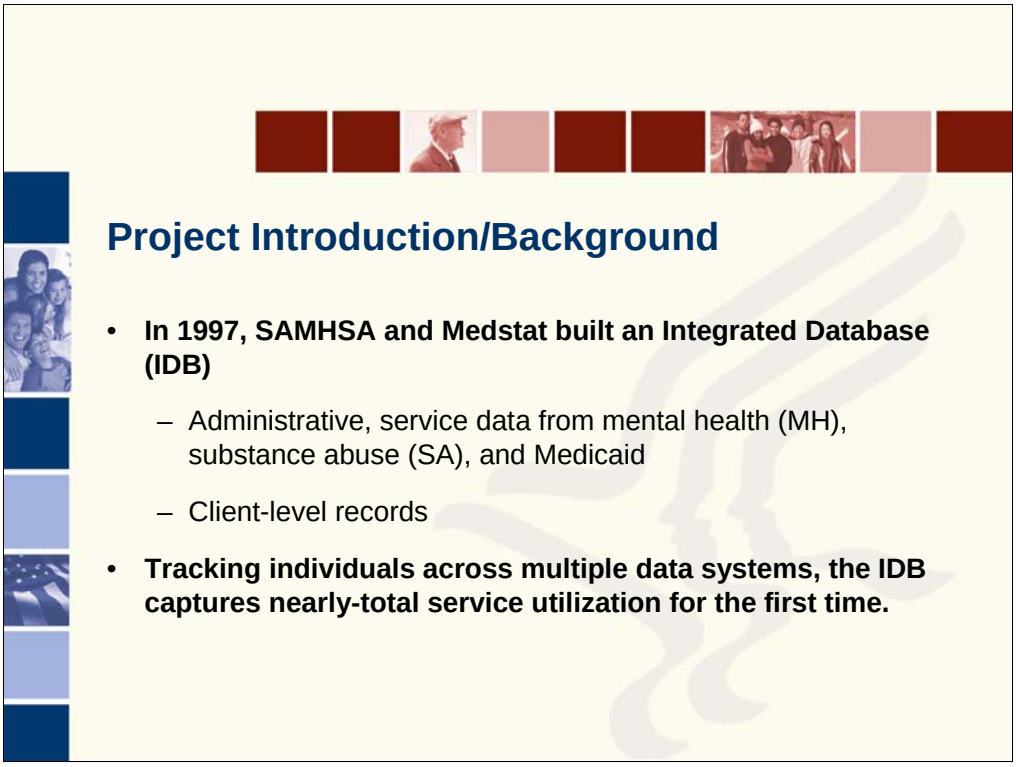
The goal is to present ideas that you can use to generate results that are useful to your state. This presentation does not propose cookie-cutter steps. Every state is different; you need to tailor these ideas to your own circumstances.

In this presentation we will:

- Highlight some of SAMHSA’s intentions for the IDB project
- Review the data integration process
- Describe the key milestones of data integration
- Discuss interim results for each of the milestones, and finally,
- Wrap things up.



Let’s start with a brief introduction to the IDB project before we get into our “Quick Start” approach.

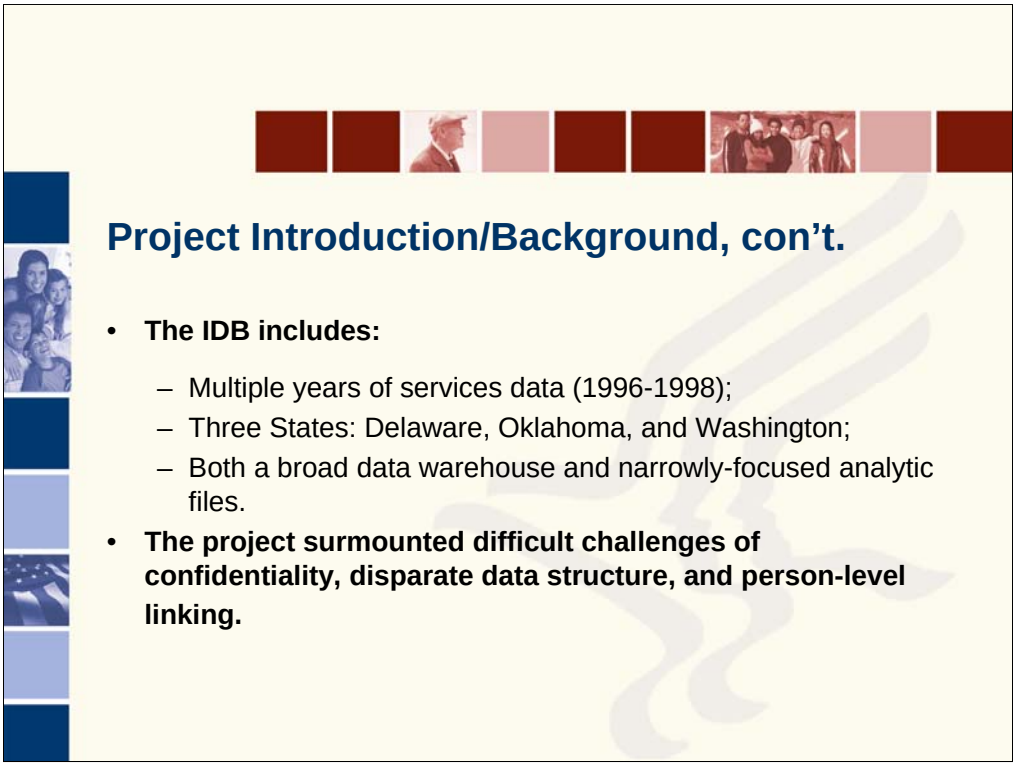


The Integrated Database – or IDB – project is a SAMHSA-sponsored contract that dates back to the late 1990s. Under the IDB project, Medstat developed probabilistic record linking techniques to integrate administrative client and service data from mental health, substance abuse, and Medicaid.

Thomson Medstat is the prime contractor for this contract. Our subcontractors include:

- RTI International
- NASADAD – the National Association of State Alcohol/Drug Abuse Directors, and
- NASMHPD Research Institute – the research branch of the National Association of State Mental Health Program Directors.

At SAMHSA's direction, Medstat created integrated databases to look at mental health and substance abuse services across agency boundaries. By gathering services from a number of sources, the IDB made it possible to more fully examine services used by people with mental health and/or substance abuse problems. A number of studies have used the IDB to examine usage patterns for people across time and service settings.



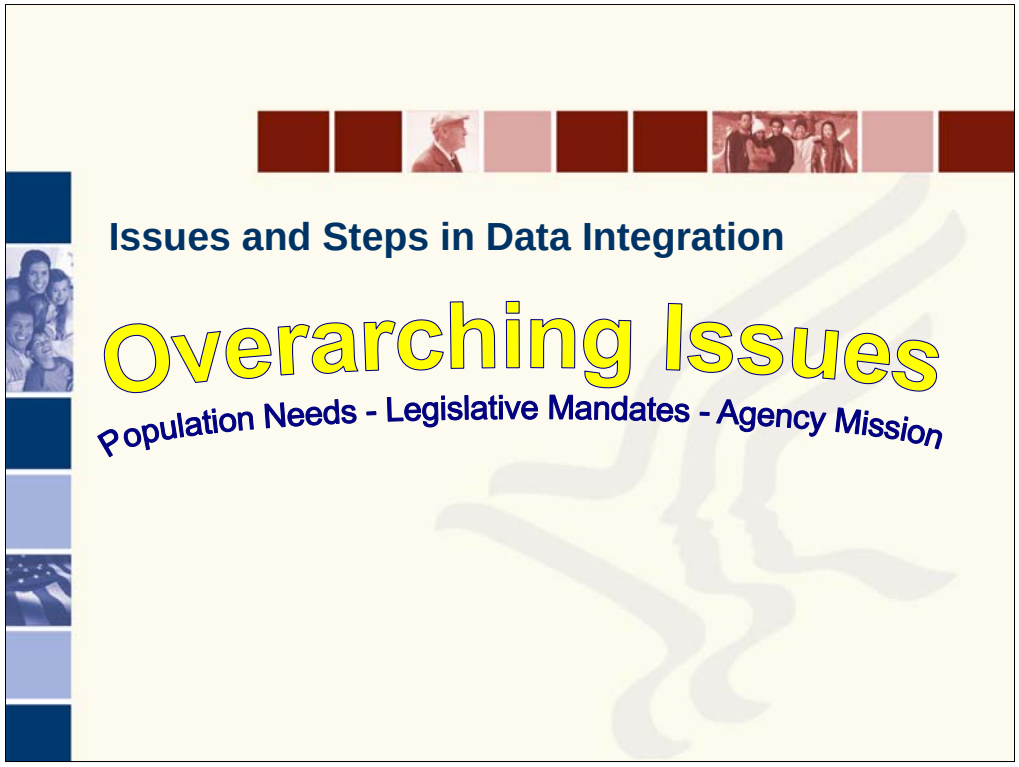
Under the previous SAMHSA contract, Medstat integrated three years of mental health, substance abuse, and Medicaid data (1996 through 1998) for three states: Delaware, Oklahoma, and Washington.

Working with these three states to build and use the IDB, we confronted numerous issues and challenges, and learned a great deal about integrating data and the issues that surround data integration projects. These include issues related to:

- Constructing an integrated database from very different data sources
- Using integrated data, and
- Maintaining confidentiality.

Based on these lessons, we have devised an approach to working through a data integration project. One of the goals of this project is to share the knowledge we gained in building and using the IDB – this presentation is part of that effort.

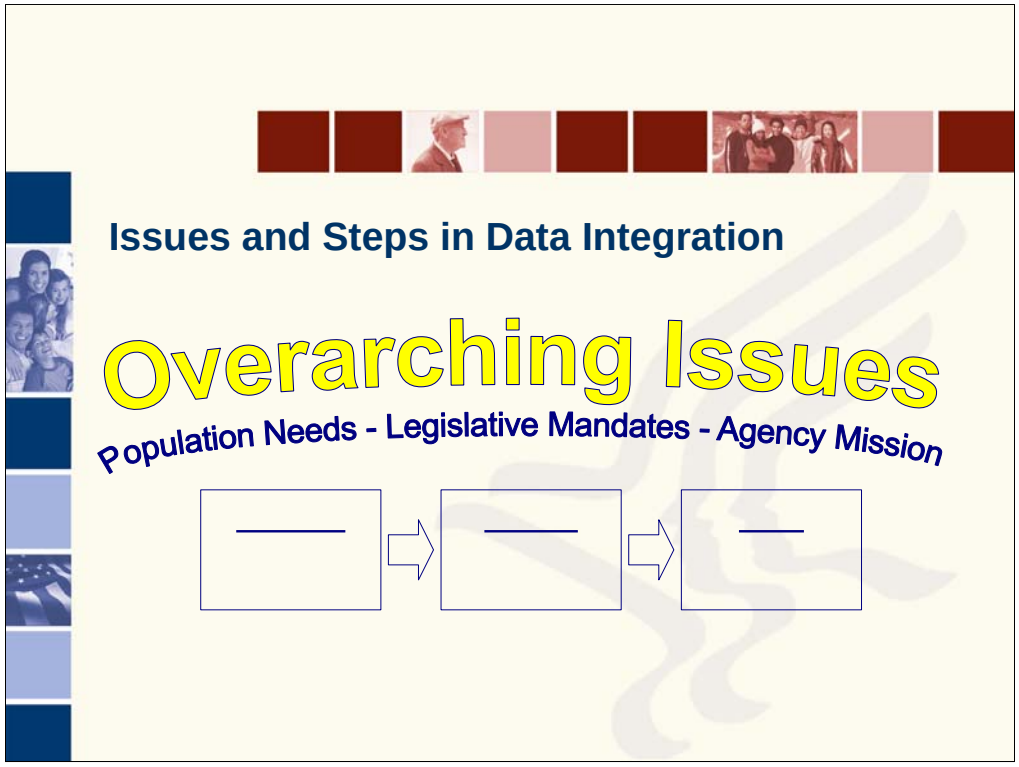
There is a lot involved in creating an integrated database. Building the database is only part of the process. Let’s spend a few minutes discussing data integration.



Data integration involves more issues than we can discuss in one presentation. Today we are focusing on an *approach* to the building process.

This slide is designed to put the building stage in some context. As with any project, there are two levels of questions to ask at the outset of a data integration process. The first layer reflects the environment in which you operate. Here, strategic questions arise, such as:

- What information is needed, given the State's population?
- What legislative mandates must be addressed?
- What is the agency's mission?



The second layer of questions deal with logistical issues such as planning and building decisions, as well as questions about how to use the integrated data after it is available.

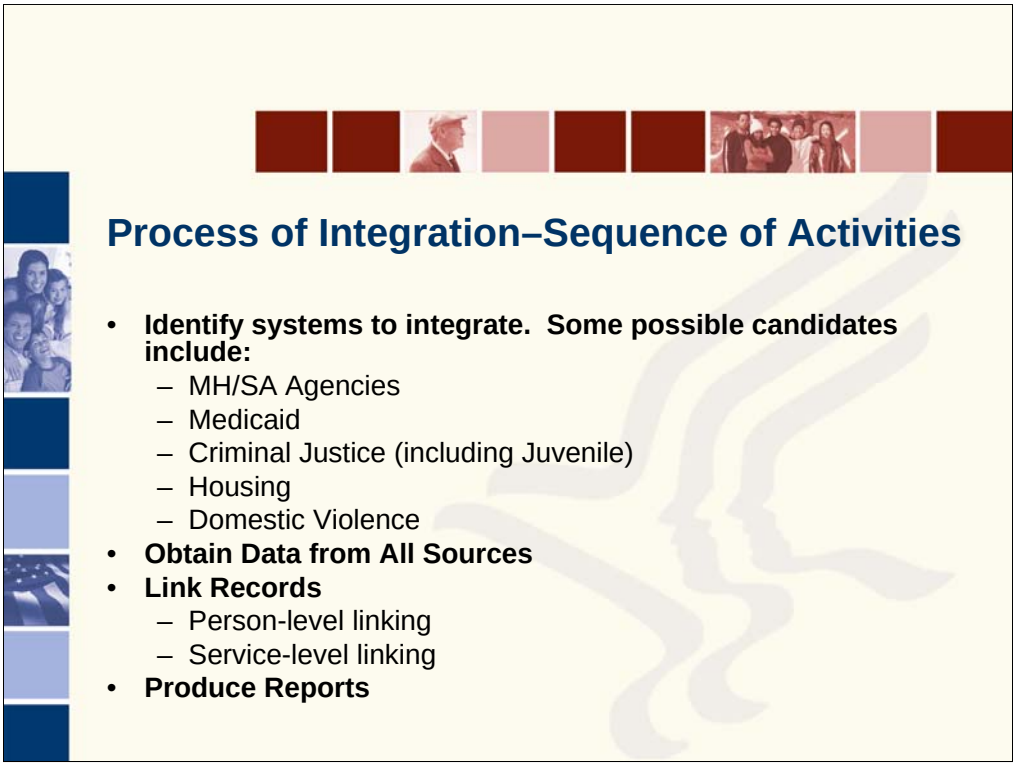
Initial steps before beginning an integration project include decisions about what to measure, whom to track, and how often to report.

There are resource issues that must be resolved, such as personnel, hardware, and software.

These are all very important – they need to be answered before deciding which data to integrate – but they are not the focus of this presentation. Once again, **this presentation addresses the "building" stage of data integration.**

Preparing
Resources
Decisions

Bu
Agre
Pro



The first step in the data integration build is identifying data that can answer your compelling issues. What are some likely issues? Mental health and substance abuse agencies are naturally concerned about patients with co-occurring mental health and substance abuse disorders. That is one example of areas to integrate.

In another example, the MH/SA Agency and Medicaid have operational needs to integrate data if:

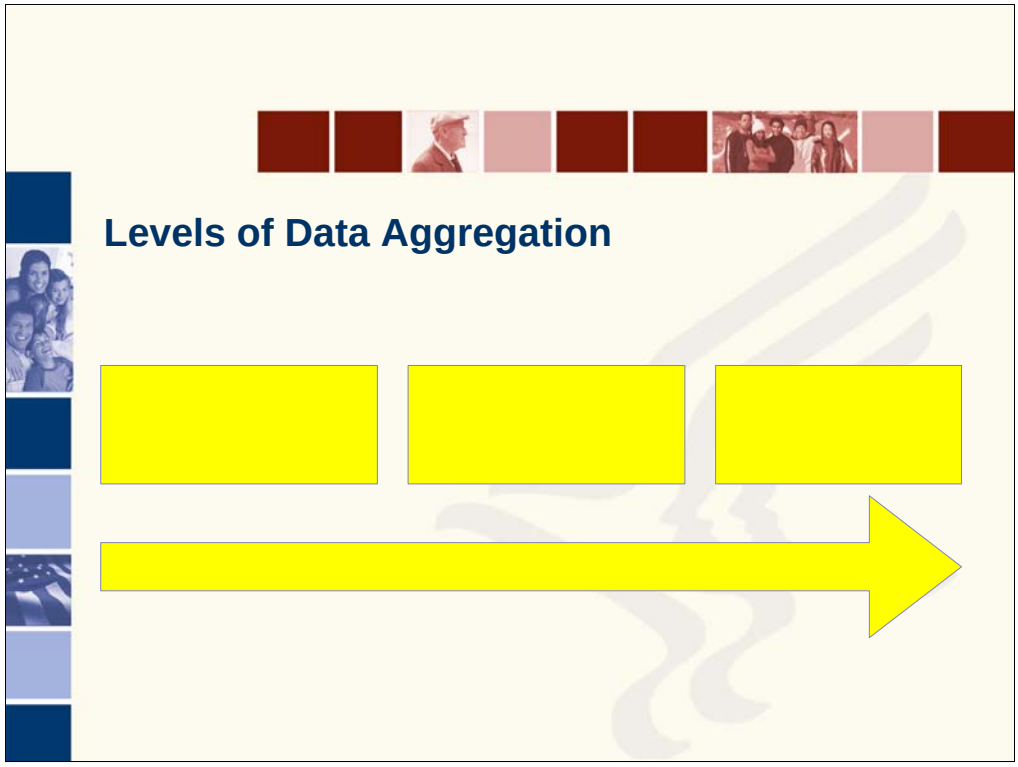
- The Agency was interested in the Medicaid prescription drug usage of MH/SA clients, and
- Medicaid was interested in high-cost users of behavioral health drugs.

Overlapping concerns may be found between MH and Employment/Labor, or between SA and Corrections. But the data you use will depend on needs particular to your State and agency; every state is different.

This presentation is focused on mental health, substance abuse, and Medicaid data integration. However, these steps and this approach can be applied to integrating other data as well. Other systems that you might consider include: Criminal Justice, Vital Records, Employment, and Education data.

The next step is to acquire that data. Allow ample time for this, as there are many issues that can slow things down.

With the files in hand, linking and then data integration can begin. And when the data is integrated, you can create ongoing reports.



As a concept, integrated data seems simple enough. But there are actually several levels to data integration.

The first level is linking *clients*

- For example, associating John Doe in one data source with John Doe in another data source.
- With client-linked data, you create a collection of service records.
- For some states, this may be all the integration possible, because your MH/SA data does not contain the detail necessary to progress further.

Client Level

Per
Elig
Enr

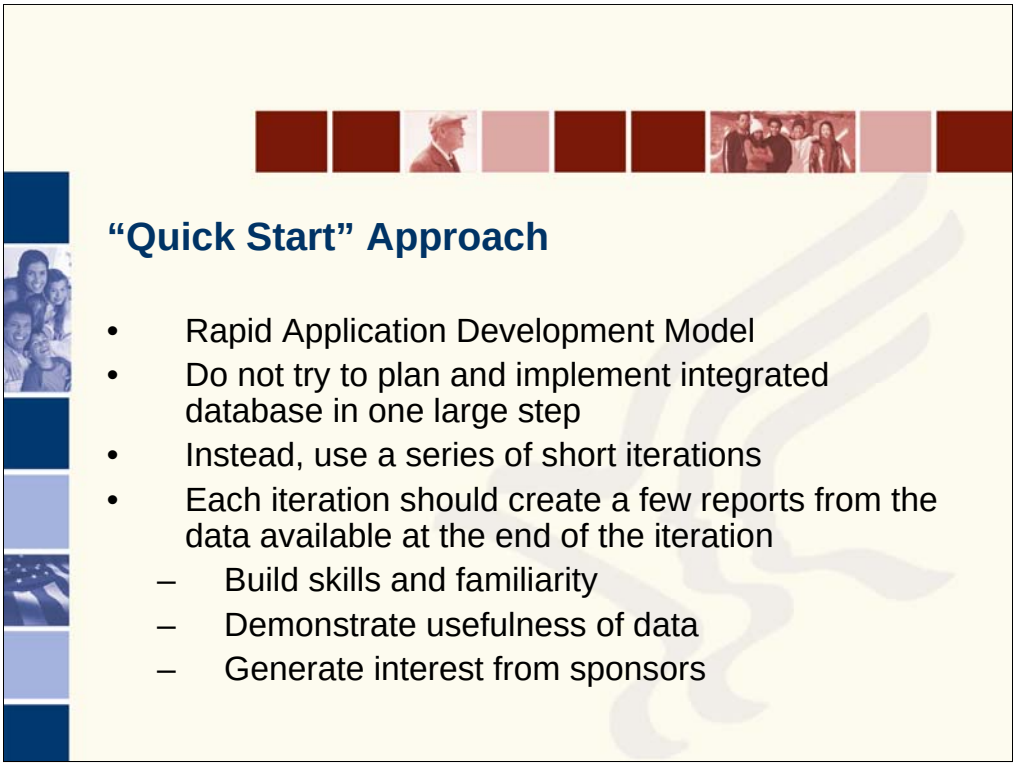
But if your data is sufficiently detailed, full integration is possible. To create a fully integrated database, you need to look at:

- Aggregation of eligibility or enrollment periods
- Linking of provider information, and
- Integration of service records

Less Integrated

The “Quick Start” process we lay out takes into account these different integration levels and can guide you through the different steps.

Now we get to the meat of the presentation: "Quick Start" Interim Results.



What is our "Quick Start" approach?

Let's start with what it is not. It is not a quick method of integrating data; data integration takes time even under the best of circumstances. It is, however, a quick method for producing useful results.

Nor is it a one-shot process. Rather than start a data integration project and wait until a fully integrated database is available before creating reports, we suggest an iterative process.

Our goal with data integration is to build and maintain momentum: to keep stakeholders interested, solicit their comments and advice.

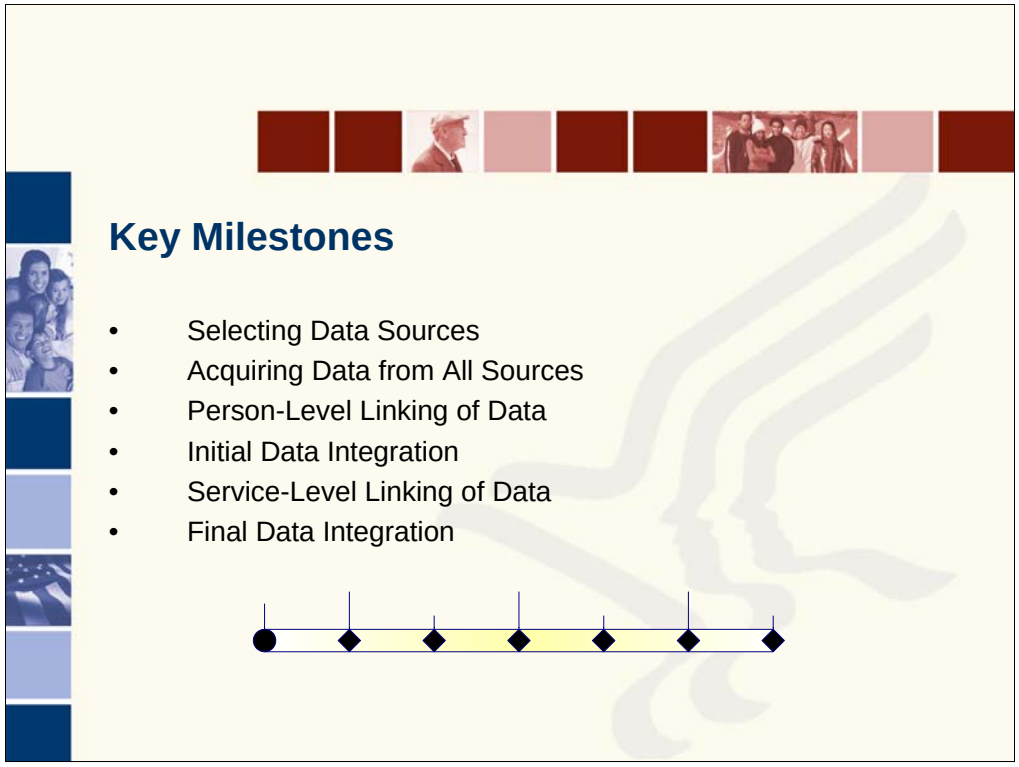
In particular, we want to hold the interest of influential stakeholders, such as agency directors and commissioners. Their feedback is important.

To do this, we suggest a multi-stage process that regularly creates reports that engage project stakeholders: a process similar to Rapid Applications Development. The process includes:

- Numerous, small steps
- A tangible product – a report or an analysis – produced after each step, something that demonstrates the usefulness of integrated data and stimulates interest in moving to the next step.

Dividing data integration into a series of smaller steps leads us directly into milestones.

By defining key milestones in the data integration process, we can create natural division points for data integration. Let’s discuss the milestones of data integration.



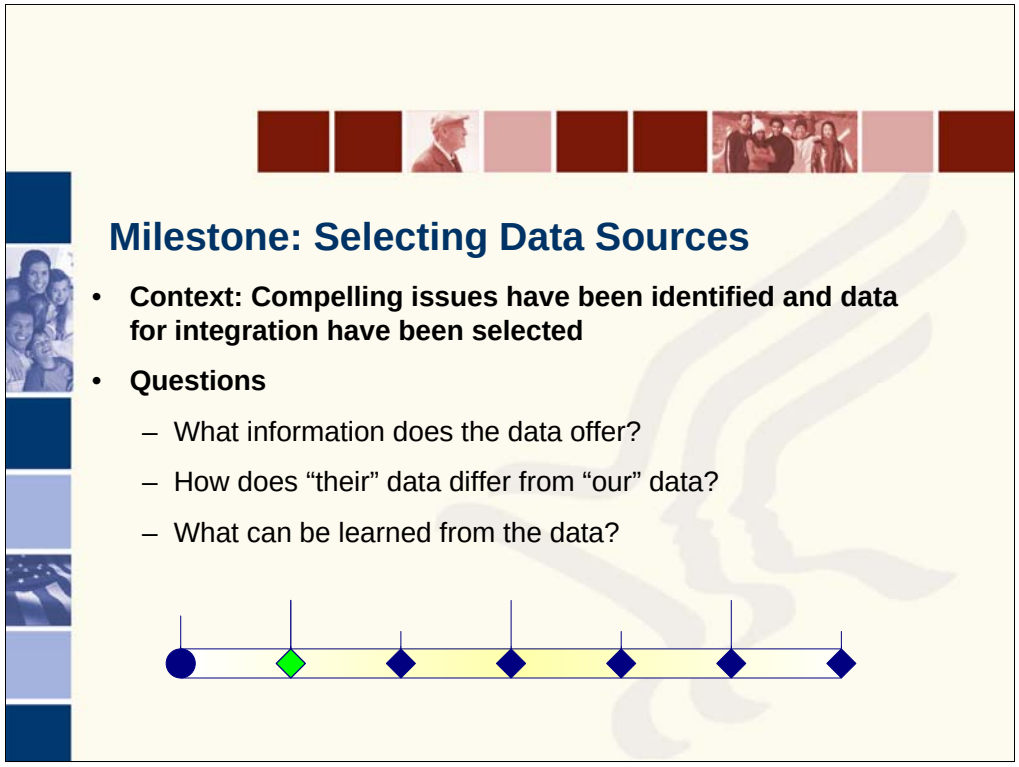
Let's identify some critical milestones.

- Selecting data sources
- Acquiring data
- Linking person-level data
- Creating the initial data builds
- Linking service-level data, and
- Building the final integrated database.

These key milestones create natural division points for the process. At each step, new information is available, new information that is useful for policy and operations.

We will walk through each of these points and discuss the type of work appropriate for each. At each milestone, we will frame the circumstances and describe the types of questions that can be addressed. Then, we will offer a reporting example and discuss the purpose of that report.

Selecting Data Sources Person-Leve
Start
Acquiring Data from All Sources



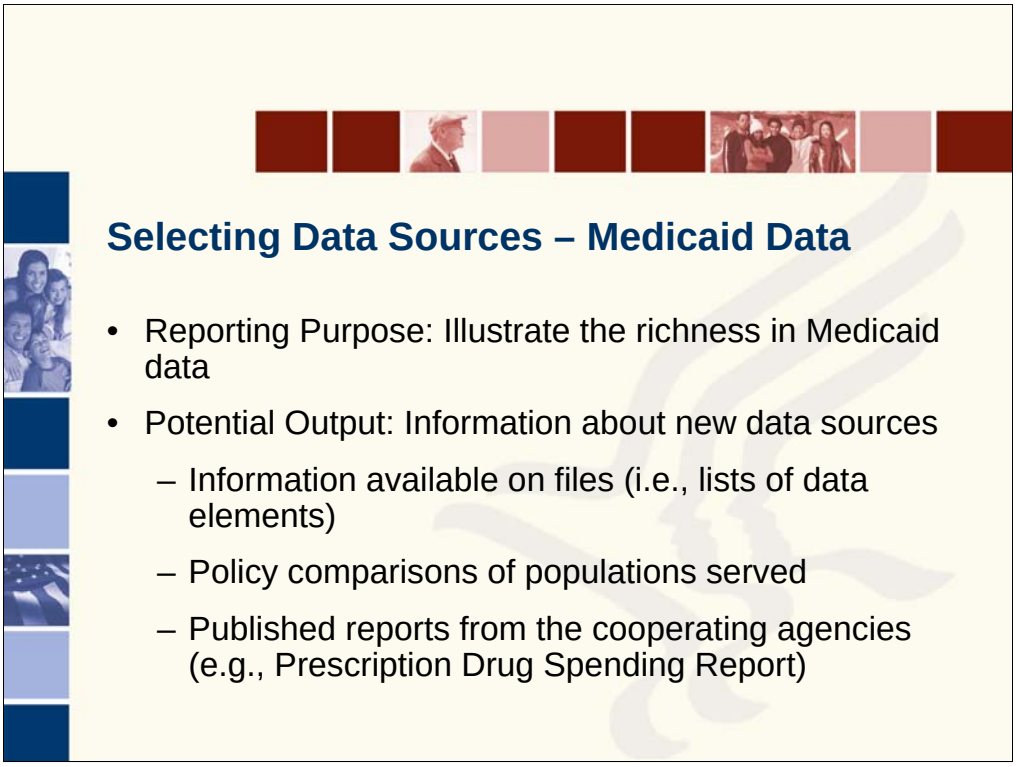
The first milestone is selecting data to integrate. We should note that selecting data sources is the *milestone*, not the first *step*. The first step is identifying the problems or questions that can be answered with more complete – integrated – data.

As noted earlier, the first step is to identify the compelling issues where different agencies need to work together. These issues drive data selection. The data you then integrate are dependent on the issues important to the cooperating agencies.

The reasons to integrate data will indicate which data to use. At this point, the data are not physically available, they have only been identified. However, meta data – information about the data – is available: information such as file layouts and data summaries. Meta data can help you determine how to use the new data once it is available.

A file layout, for example, will inform you as to what data elements to expect in the data. Comparing that information to the information on your own files demonstrates how the new data will supplement existing data. Other layout information will indicate elements, such as race fields, where cross-walks are necessary between two files.

We’ll use Medicaid data to illustrate our point.

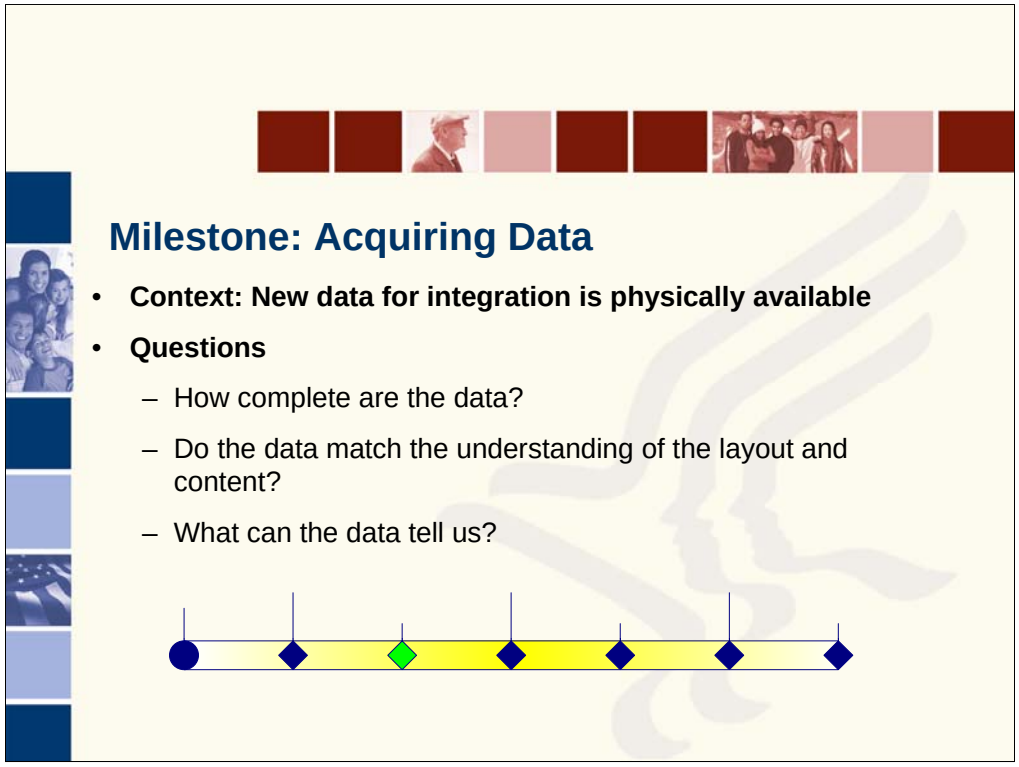


The goal at this milestone is to illustrate the usefulness of the chosen data. Your initial reports should set expectations and also reassure stakeholders that data integration is a sound decision. We can demonstrate the value of the Medicaid data and continue building interest in the project. This may require analysis and interpretation. The report should engage stakeholders and give them a chance to provide feedback.

Logical starting reports are layouts or data dictionaries for the Medicaid files. Data dictionaries do not excite many people (except for programmers), but adding context changes things. Using a list of data elements to point out the types of analyses the information makes possible can generate interest and build support for data integration projects.

Another useful report is a comparison of “eligibility” criteria and covered services. A report like this is often helpful when analyzing usage patterns in the later stages. For example, the setting of detoxification services often depends on a client’s Medicaid eligibility. Clients with Medicaid eligibility usually receive inpatient detoxification, while clients without Medicaid usually receive residential detoxification.

Published Medicaid reports are another valuable source of information. Often these can be compared to mental health and substance abuse reports.

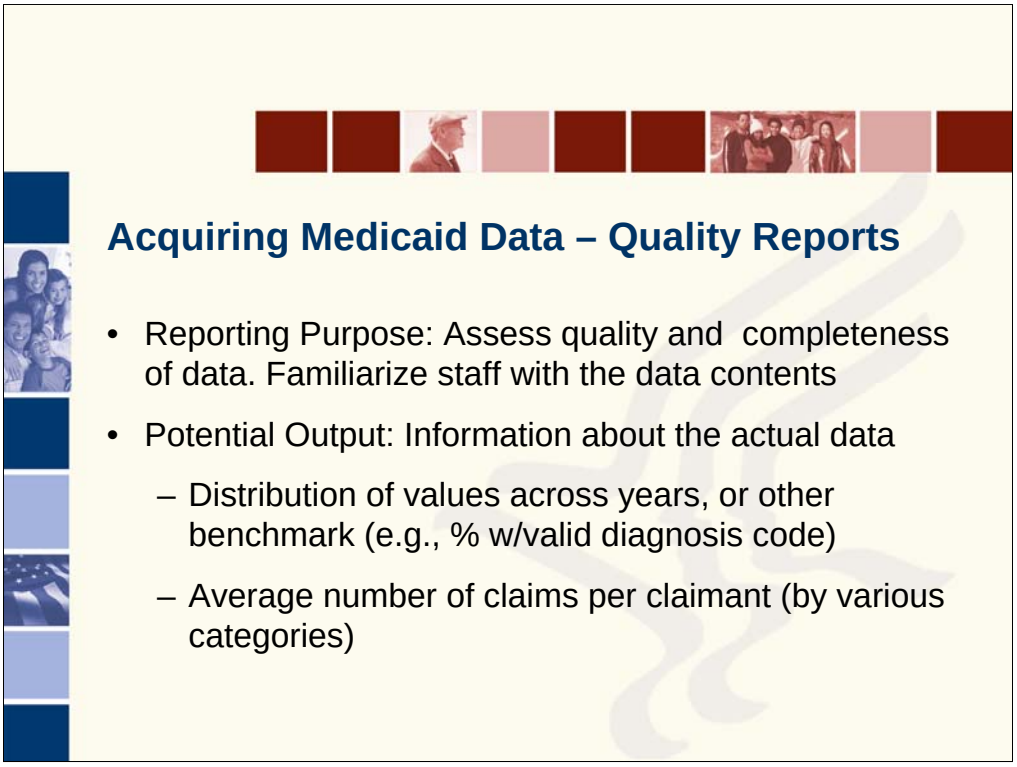


The next milestone is acquiring data. Negotiating data-sharing agreements and identifying resources may delay your acquisition beyond your wildest dreams. So it is important to be ready when the files do arrive. Familiarity with file layouts can be helpful because it will speed your work with the new data.

However, it is likely that you will not have used the new data before. So once you have the files, you need to explore, or examine the data. Look at the quality and ask:

- How complete are the data? What is missing? How often is it missing?
- Do record and client counts match reported totals? Make sure you understand the data.

Finally, start looking at what the data reveal.



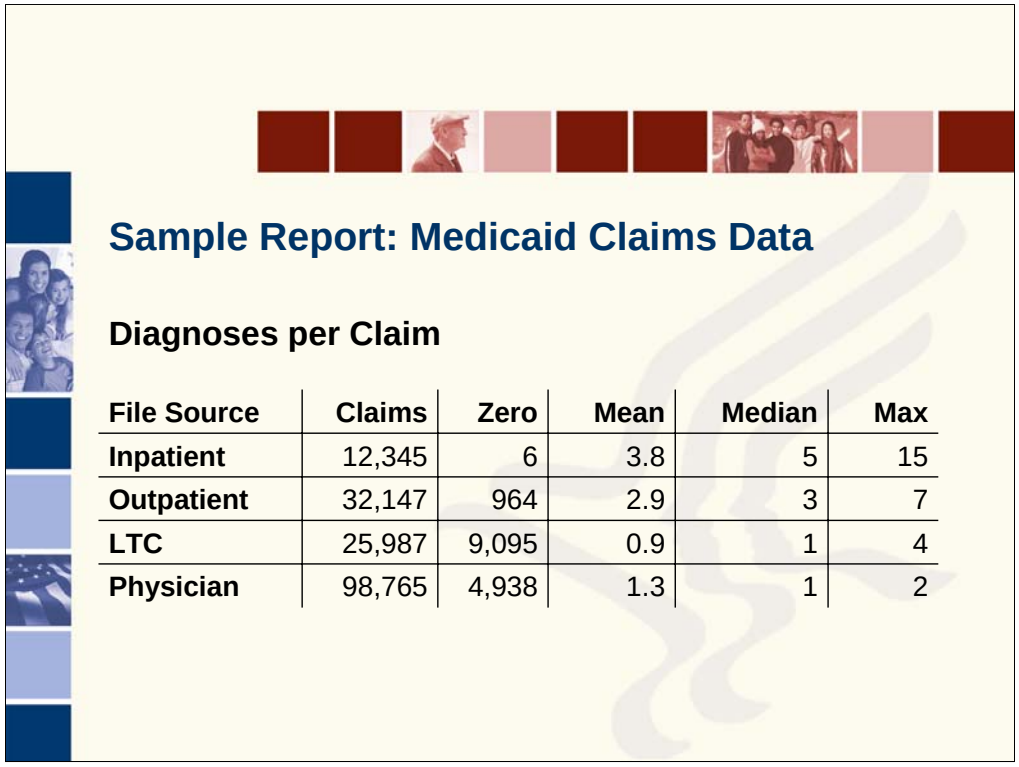
To answer those questions, you need to explore the data. Regardless of the type of data, your first step should be to assess data quality. This is an excellent way of gaining familiarity with a new data source. Some items to examine include:

- Comparing record counts expected and received
- Contrasting unique client counts between Medicaid eligibility files and the reported Agency numbers.

Because this is a new data extract, you might get bad data. However, don't jump to that conclusion until you are sure: it may be possible that you do not fully understand the data. Perhaps the file layout was interpreted incorrectly. A common mistake involves over-counting inpatient claims by not correctly handling Medicaid claim line-items – a topic that is presentation-worthy in its own right.

For Medicaid data, here are some other quality assessment reports you should consider:

- Distribution of values for specific fields (or variables), such as race or date of birth
- Averages and totals for various categories, such as amount paid
- Percentage of records with valid diagnosis codes (or specific diagnosis codes).



Here is a report that explores the completeness of Medicaid claims data by looking at how often diagnoses are coded on the claims data. Left to right, columns in this report illustrate claim file sources, the number of claims for each category, the number of claims with no diagnosis, the average number of diagnoses, the median number of diagnoses, and the maximum number of diagnoses.

This example, like most of the other examples provided in this presentation, is simplified and completely fabricated. But it does convey the type of information you might want to examine.

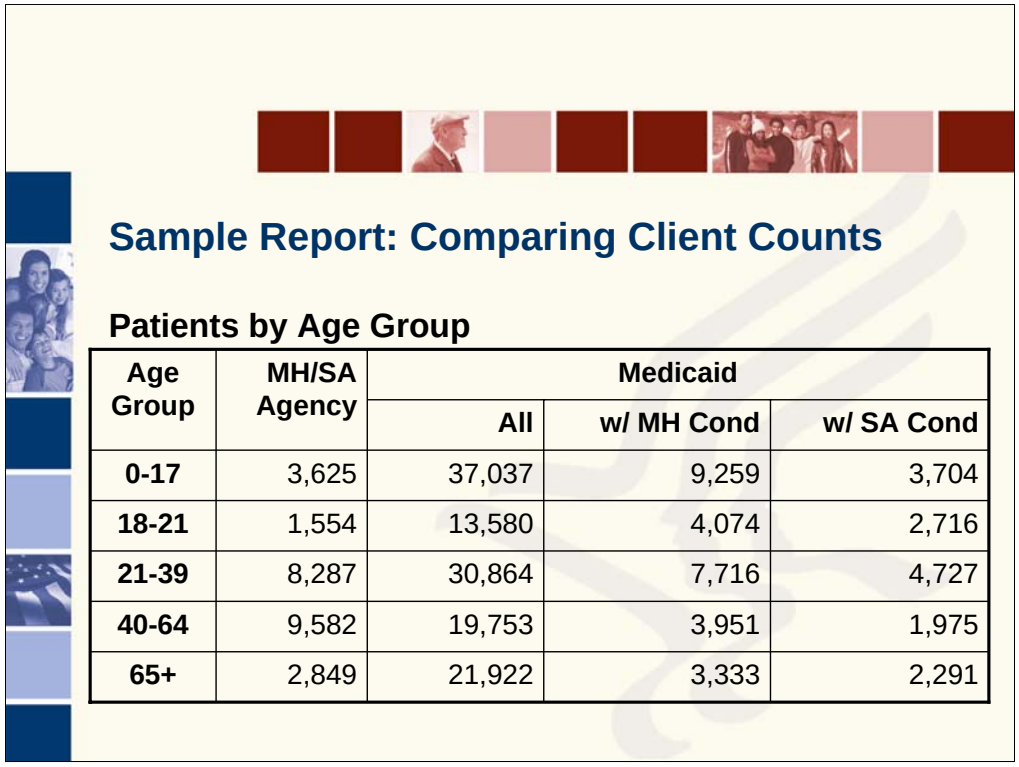
This report involves only Medicaid claims data. Remember, at this point, the data has only been received: no integration with MH/SA data has occurred.

Further steps might include: grouping diagnoses into categories; finding the diagnoses that occur most frequently; and examining co-morbidities.

With data available and examined for completeness, you can and should explore the data more deeply. Once you are somewhat familiar with the data, you can create some interesting reports that can provide information about clients with mental health or substance abuse disorders. These reports can also build project momentum and help drive the integration process forward.

You may want to examine data sources separately, breaking down usage and patient counts by types of claims or service, or by different demographic measures. You could also look at claim counts by different categories, such as claim or service type.

Another useful comparison is to investigate population characteristics for Medicaid clients versus mental health/substance abuse Agency clients…



Sample Report: Comparing Client Counts

Patients by Age Group

Age Group	MH/SA Agency	Medicaid		
		All	w/ MH Cond	w/ SA Cond
0-17	3,625	37,037	9,259	3,704
18-21	1,554	13,580	4,074	2,716
21-39	8,287	30,864	7,716	4,727
40-64	9,582	19,753	3,951	1,975
65+	2,849	21,922	3,333	2,291

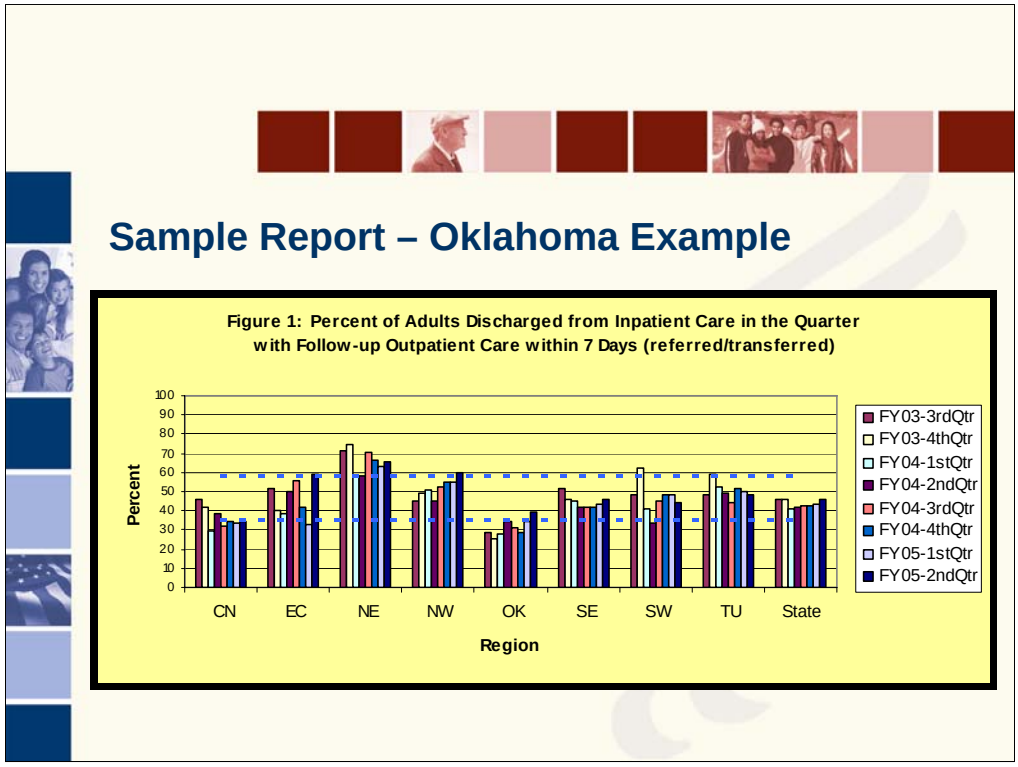
… as this sample report does. This report compares mental health/substance abuse client counts to Medicaid client counts across age-groups.

This example lines up population breakdowns to compare Agency clients with:

- All Medicaid patients
- Medicaid patients with a mental health condition
- Medicaid patients with a substance abuse condition.

An informative modification to this report would be to add percentages; we have not done so here because of space restraints.

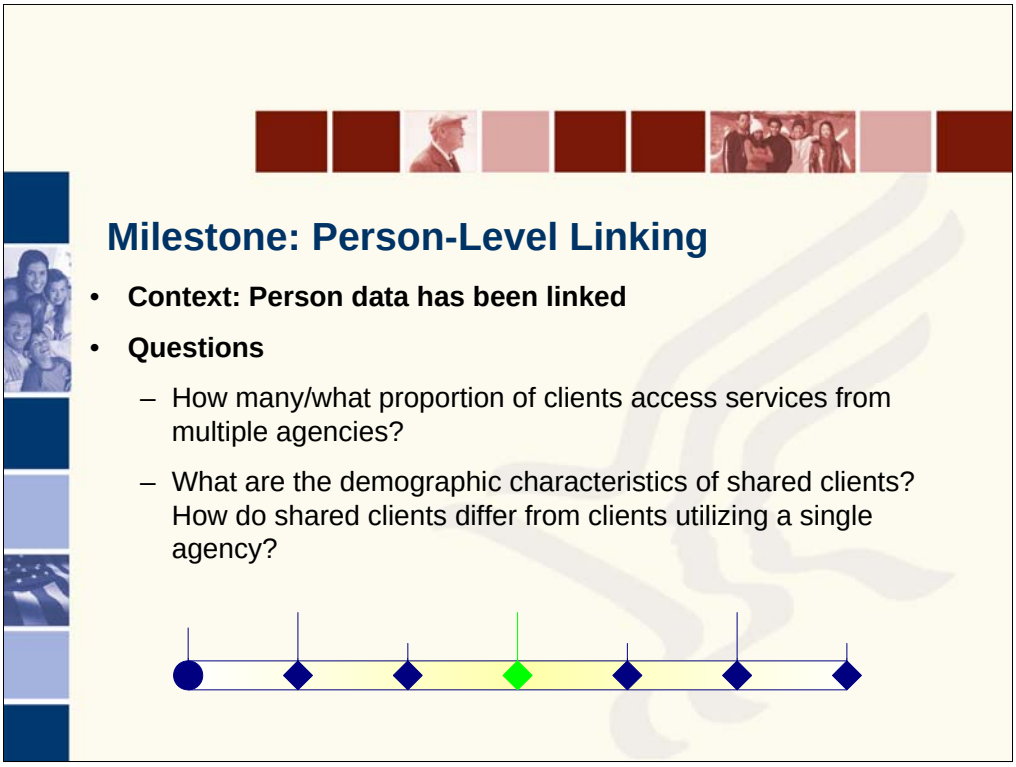
These numbers are revealing, but they do not tell the whole story. This report looks at MH/SA and Medicaid data separately. The question you should ask rhetorically is, “What is the patient overlap between the two agencies?” Obviously, this question cannot be answered until the data is integrated.



You can also look at usage over time.

Here is an example we extracted from Oklahoma's "Regional Performance Management, or RPM report, that we found interesting and useful. This graph examines follow-up after inpatient care and displays trends over the past two years. The chart shows quarterly follow-up rates. From left to right, the bar-groupings show quarterly rates for the eight regions within Oklahoma; follow-up rates for the entire State appear as the right-most grouping. The horizontal dashed lines represent one standard deviation above and below the mean.

This example uses only MH data, so questions remain. For example, how would these rates change if you included Medicaid services? To answer that question, you need to integrate the data!

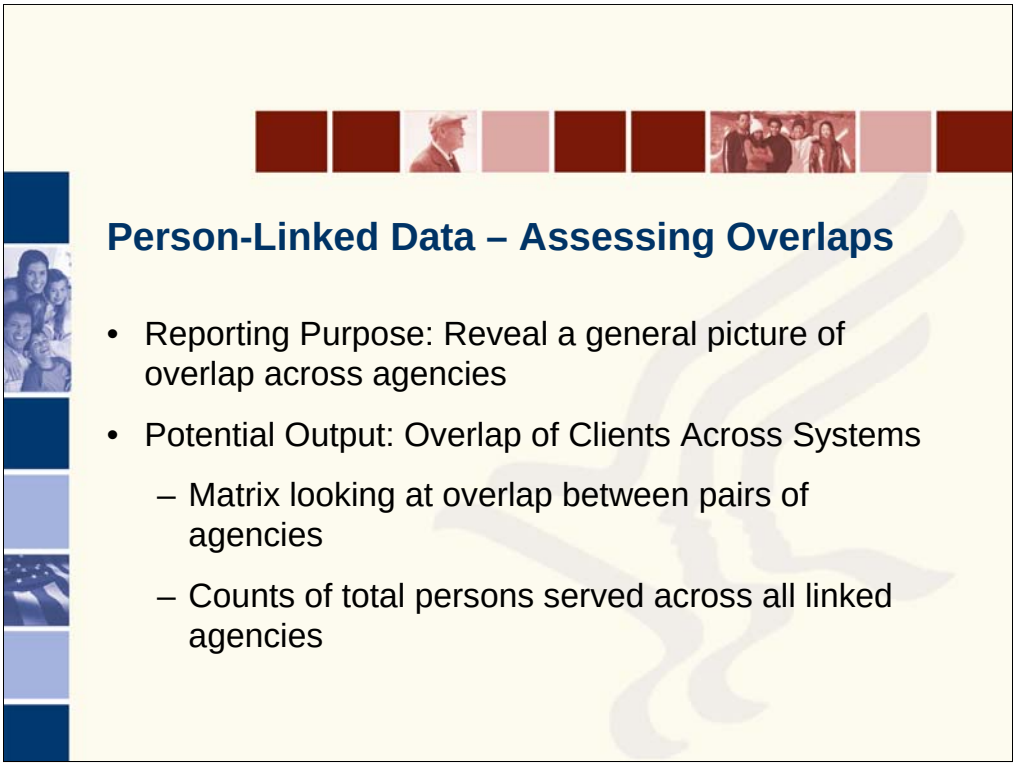


The next milestone is reached when the data have been linked at the client level.

This is a big accomplishment: linked data. Finally there is something that people have been expecting. At this point, states can demonstrate some success with the project and at the same time generate some more useful analyses. With the person-level linked data available, you can explore person-level overlaps:

- Which agencies share clients – how many clients and what proportions?
- What is the total number of served clients across all systems?

You can also examine the demographics of the client population – comparing characteristics of shared clients and non-shared clients.



This is a point where you can learn a great deal about the underlying demographics of the behavioral health clients. In our examples, behavioral health clients are people who receive services through mental health/substance abuse agency funding and/or through Medicaid coverage.

It is important to remember that at this stage we are talking about client-linked data, so it makes sense to focus on client information. Using integrated mental health, substance abuse, and Medicaid data and contrasting population demographics, you can examine how clients are similar or how they differ. The types of groupings to consider include:

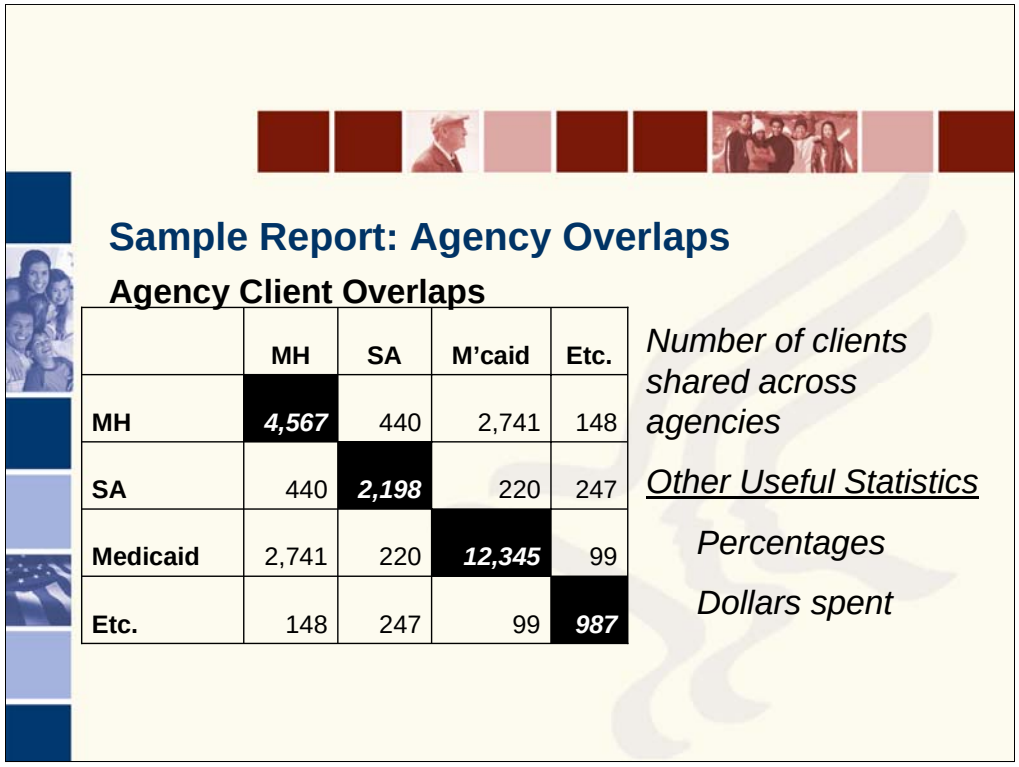
• The entire population serviced through the mental health/substance abuse agency

• The entire population serviced through Medicaid

• The population serviced through both mental health/substance abuse and Medicaid – in other words, the overlap

• The population serviced only through the mental health/substance abuse agency – the mental health/substance abuse agency population less the overlap

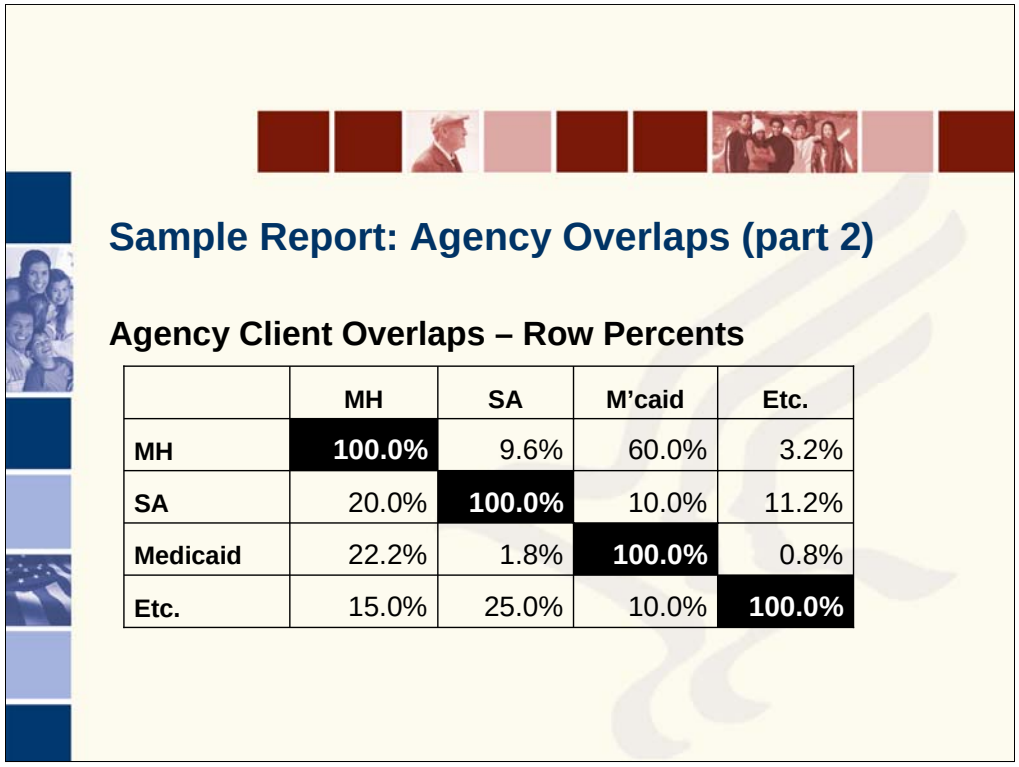
• The population serviced only through Medicaid – the Medicaid population less the overlap.



Our example for this milestone illustrates overlaps between agencies. This is a layout that we copied from work done in the State of Washington. The format is quite flexible and can expand or contract to accommodate several agencies.

The shaded diagonal represents the client population for each agency – in this example – of a fictional population – mental health has 4,567 clients, substance abuse has 2,198 clients, and so on. Other cells represent the clients shared by the agencies in the column and row headers – mental health and substance abuse share 440 clients; substance abuse and Medicaid share 220 clients.

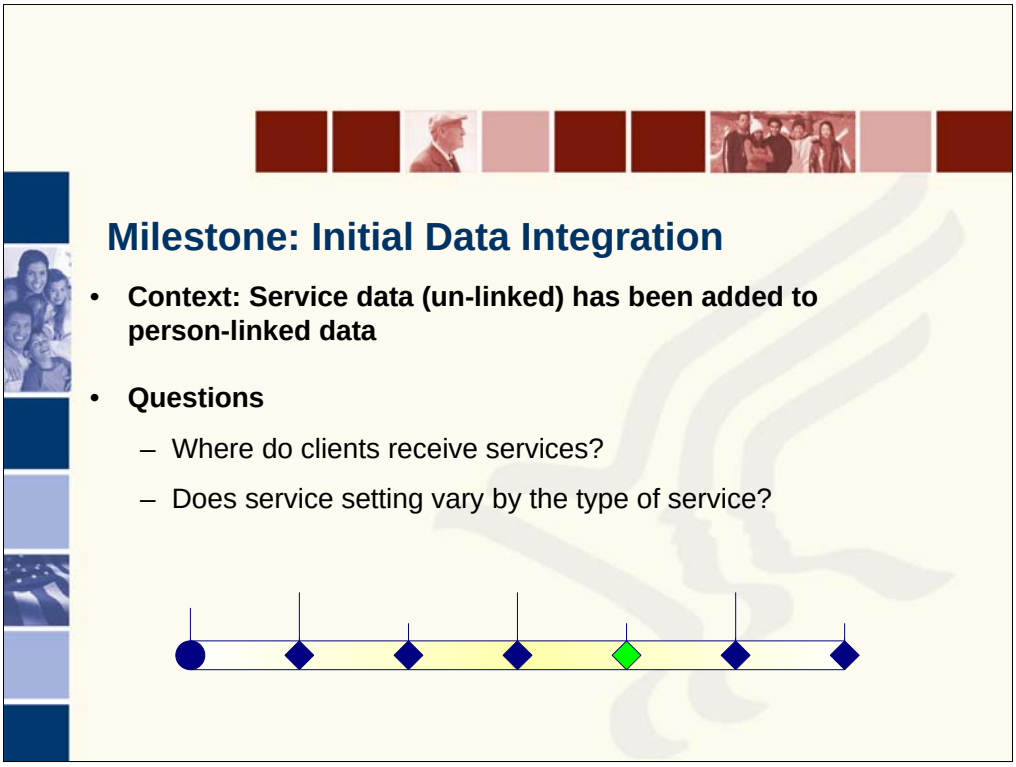
This example presents agency overlaps in raw numbers. Percentages would be a useful addition to this report to show the relative number of shared clients.



That information is demonstrated in this table, where the diagonals total 100%. The column-agency is the numerator and the row-agency is the denominator. The table is no longer symmetric.

Another approach is to look at where services are tracked. By including the pooled services – remember they are not unique services because there has been no service-level linking – you can also see who receives services through which agency.

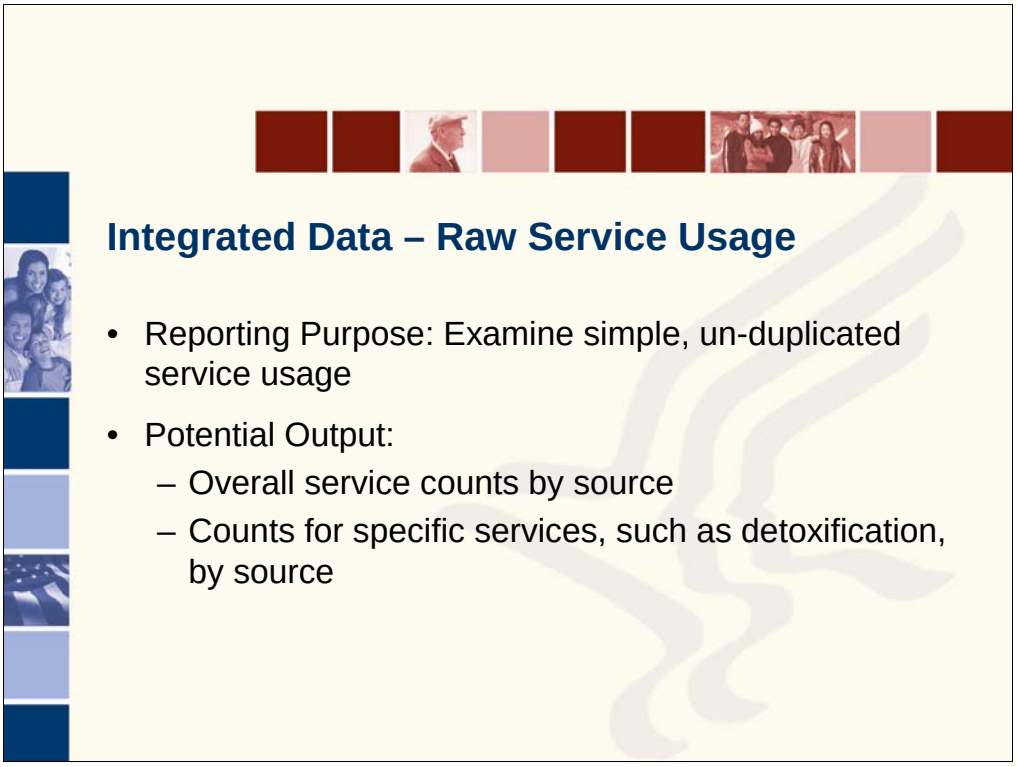
In these examples, you do not know total service usage because you have not yet linked at the service level.



With person-linking complete, it is possible to look at services. This is the initial integrated database. At this point, only clients have been linked; no service-level linking has taken place. So there is the potential for duplication of services between the various source data. For that reason, we do not consider the data fully integrated.

However, it is possible to generate reports that are informative, useful, and interesting at this stage. The raw, un-duplicated services can still inform policy. Analysis at this stage can reveal information about how and where clients receive behavioral health services.

Reports at the previous milestone focused on person-level information. In contrast, the emphasis at this milestone is service-linked data.



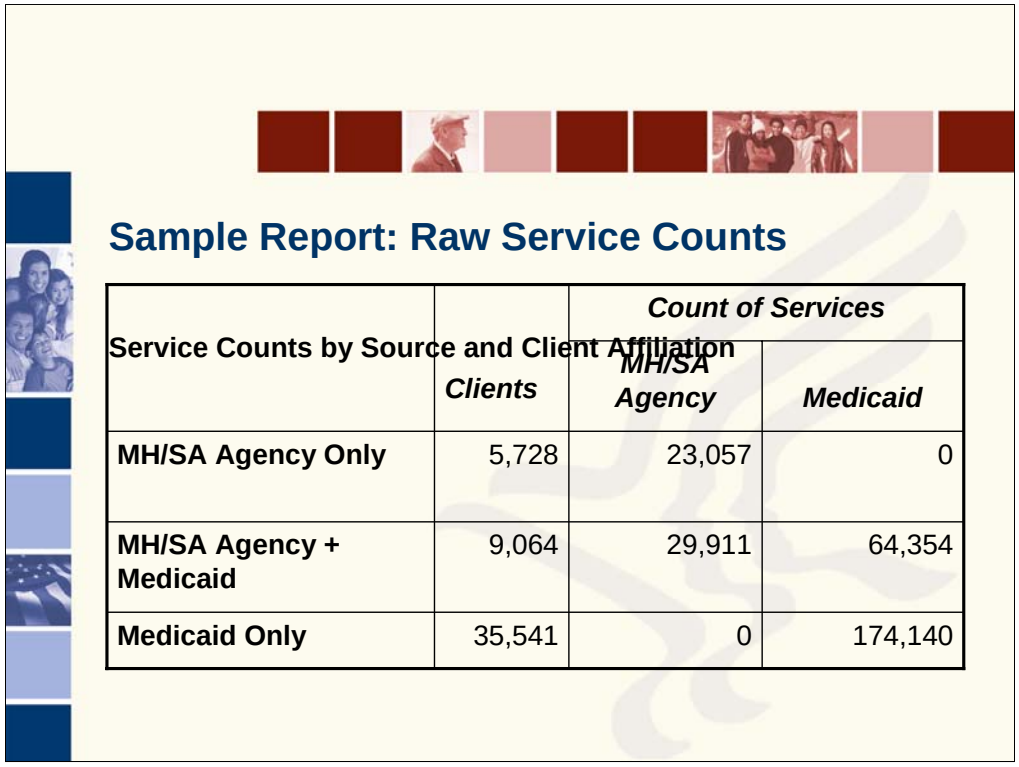
Simple, unduplicated services can answer these questions:

• Which services do clients use?

• Who provides their care?

One example of potential reporting topics is to illustrate breakdowns of service counts by servicing agency. You can do this for various client groupings, such as demographics or the agencies through which people receive services.

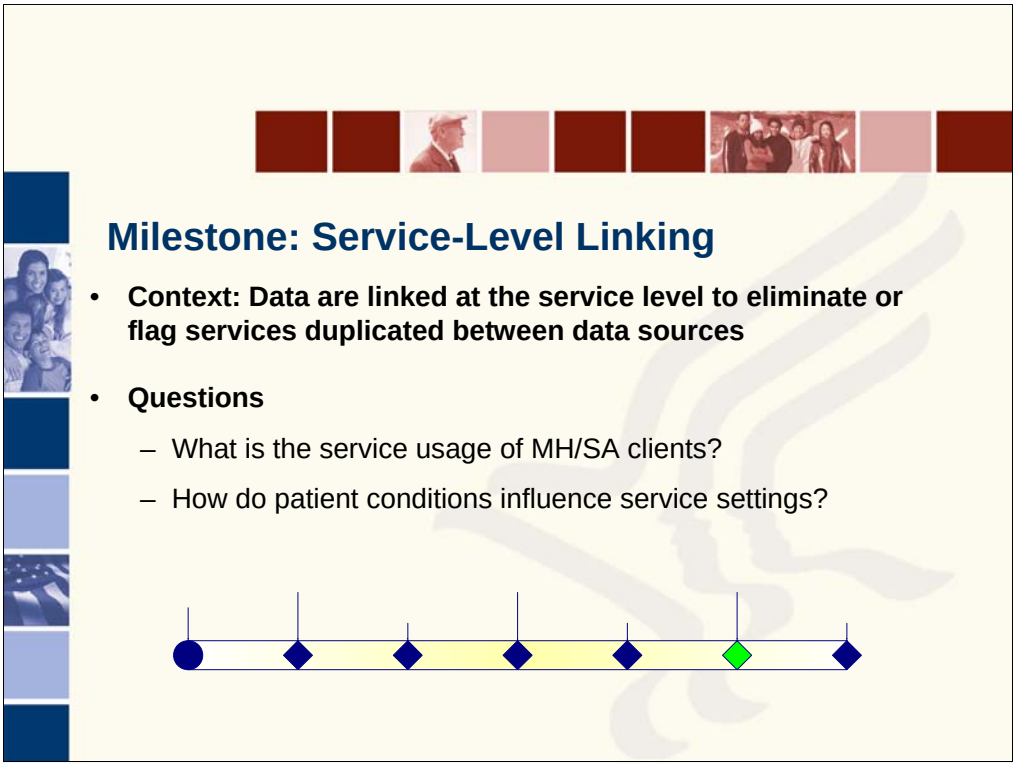
Other examples include counts of specific service types, again for various client groupings.



This sample report illustrates a simple service count report. A useful addition would include average services per client, and perhaps even quartiles. In addition, you may wish to add combined groupings, such as all mental health/substance abuse agencies and all Medicaid.

Let's indulge in a brief tangent. Our examples integrate mental health/substance abuse data with Medicaid data: data with service records. Not all data are service oriented, and not all service data contain service records that might duplicate mental health or substance abuse services. Corrections data is an example. Integrating mental health/substance abuse data with corrections data does not require service-level linking. In such a case, the data are complete after mental health/substance abuse services and correction records are added to the person-linked data.

Medicaid data, however, is service oriented, and to fully integrate Medicaid and mental health/substance abuse data, it is necessary to perform some service-level linking.



Recall that person-level linking connects client records across agency systems. With person-level linking, you can associate John Smith who received services through the mental health/substance abuse agency with John Smith who received services through Medicaid coverage.

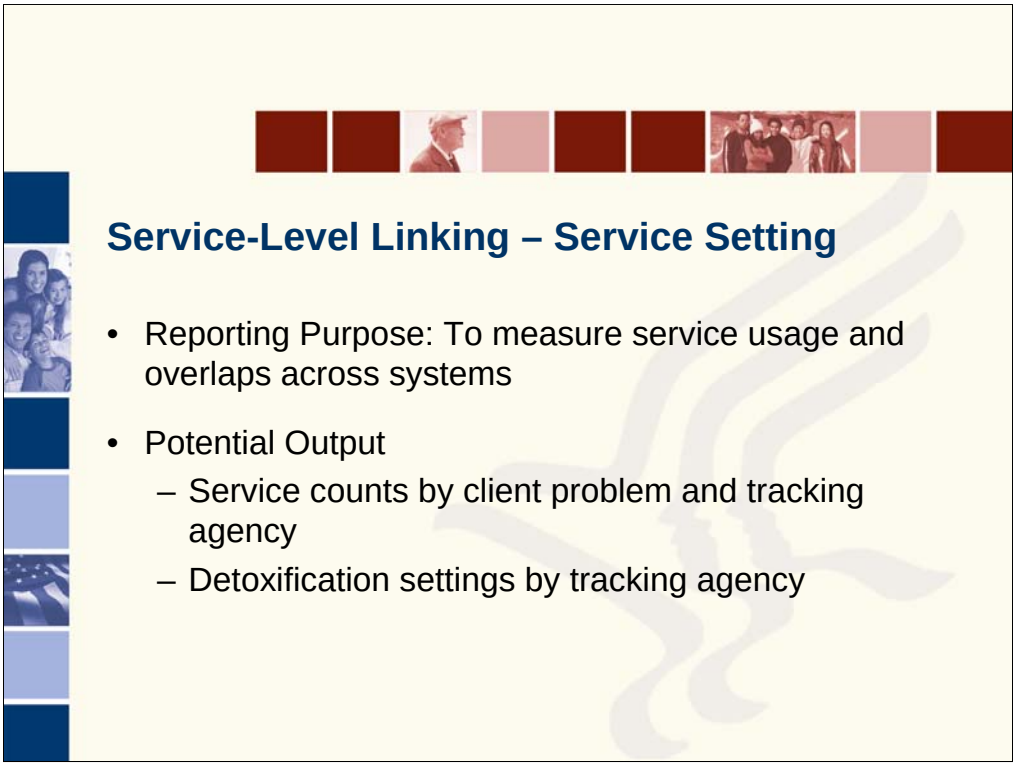
By “service-level linking” we’re referring to finding the same service in different sources. For example, mapping a mental health agency service on a particular date with a Medicaid claim for one hour of counseling on the same date. Because of limitations within many MH/SA data systems, service detail is often not available, so this level of linking is not always possible. But where it is possible, the most useful of reports can be created.

Examples of the work you can do at this milestone include looking at overall usage by the MH/SA clients and tracking episodes of care where clients receive different services through different agencies.

Selecting Data Sources Person-Level

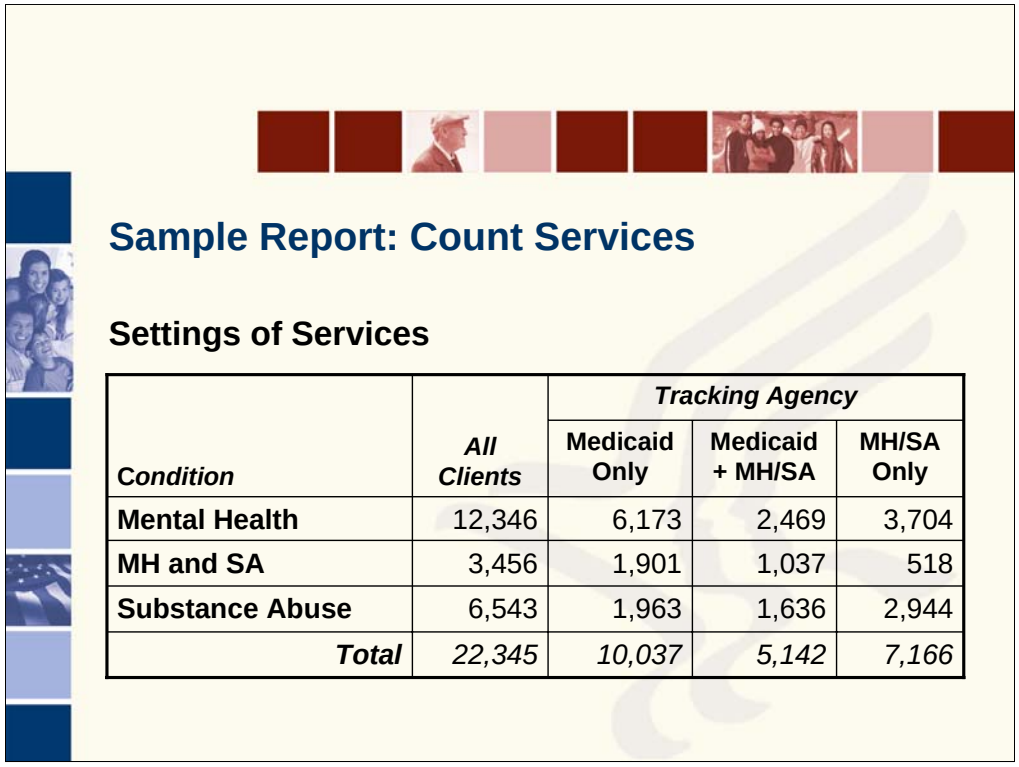
Start Acquiring Data from All Sources

With service-linked data, you can generate a more holistic view of client service usage. The example we provide for this stage focuses on service settings – which agencies track services – for different patient conditions.



Reporting service counts by client problem and tracking agency is the example we use to illustrate this milestone. The purpose of this report is to examine service settings across systems – MH/SA and Medicaid, in this case.

Now, lets turn to the report table.



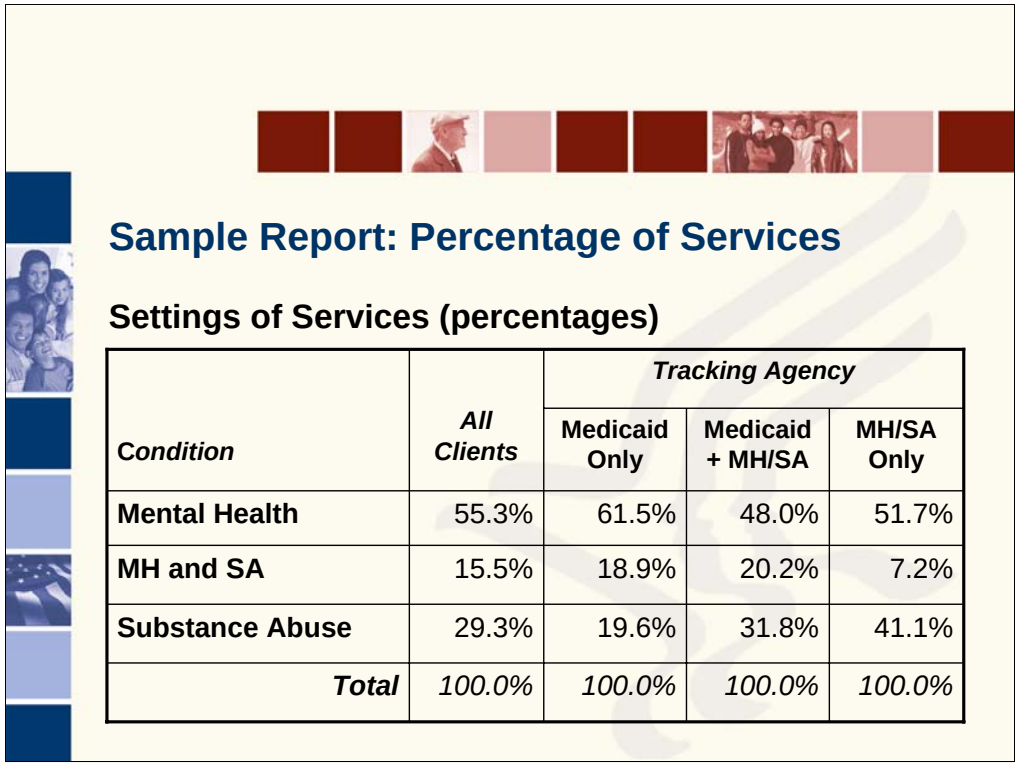
Sample Report: Count Services

Settings of Services

Condition	*All Clients*	*Tracking Agency*		
		Medicaid Only	**Medicaid + MH/SA**	**MH/SA Only**
Mental Health	12,346	6,173	2,469	3,704
MH and SA	3,456	1,901	1,037	518
Substance Abuse	6,543	1,963	1,636	2,944
Total	*22,345*	*10,037*	*5,142*	*7,166*

Some clients with mental health and substance abuse problems may receive their mental health services exclusively through Medicaid and their substance abuse services through the MH/SA agency. Examining service usage through a single source therefore fails to provide a complete picture of co-occurring clients. By integrating the two data sources, you can identify which clients have both mental health and substance abuse conditions.

In this example, we focus on people with diagnoses for mental health or substance abuse conditions or both mental health and substance abuse conditions – but not necessarily at the same time – and we look at agencies through which they receive services. This is another example that has been simplified for readability: it contains only raw (albeit made-up) numbers. A useful addition would be to add percentages.



Sample Report: Percentage of Services

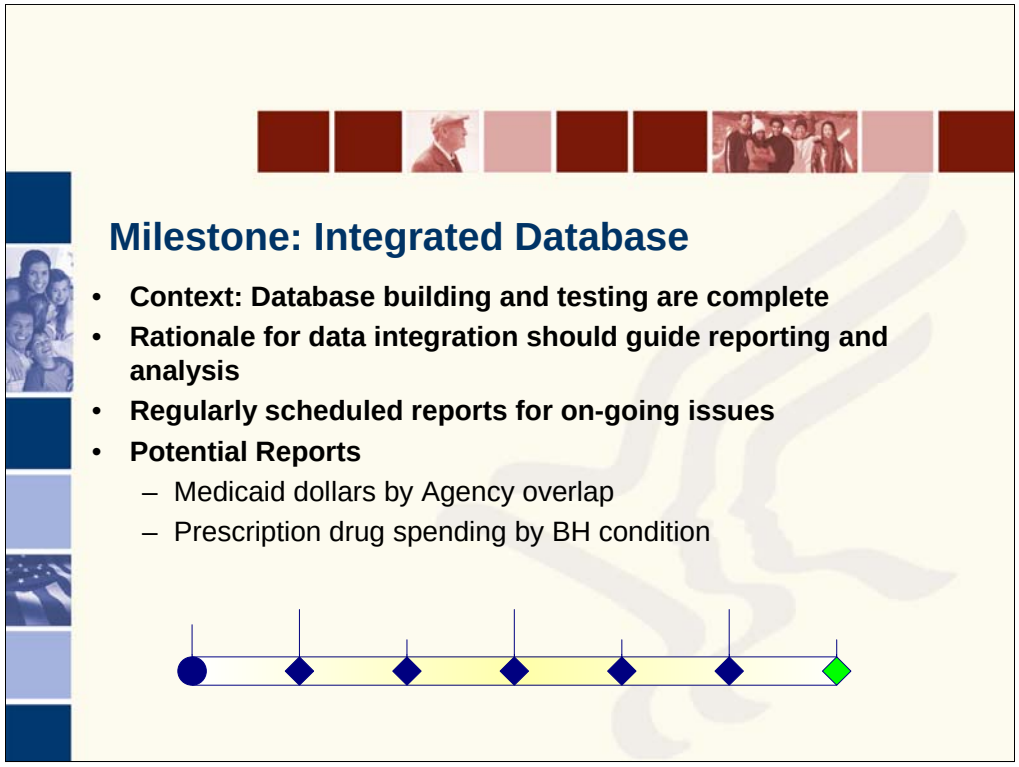
Settings of Services (percentages)

Condition	*All Clients*	*Tracking Agency*		
		Medicaid Only	**Medicaid + MH/SA**	**MH/SA Only**
Mental Health	55.3%	61.5%	48.0%	51.7%
MH and SA	15.5%	18.9%	20.2%	7.2%
Substance Abuse	29.3%	19.6%	31.8%	41.1%
Total	*100.0%*	*100.0%*	*100.0%*	*100.0%*

Something like this…

Here is the previous table expressed as column percents.

This, of course, would be much more effective when combined with the previous slide.

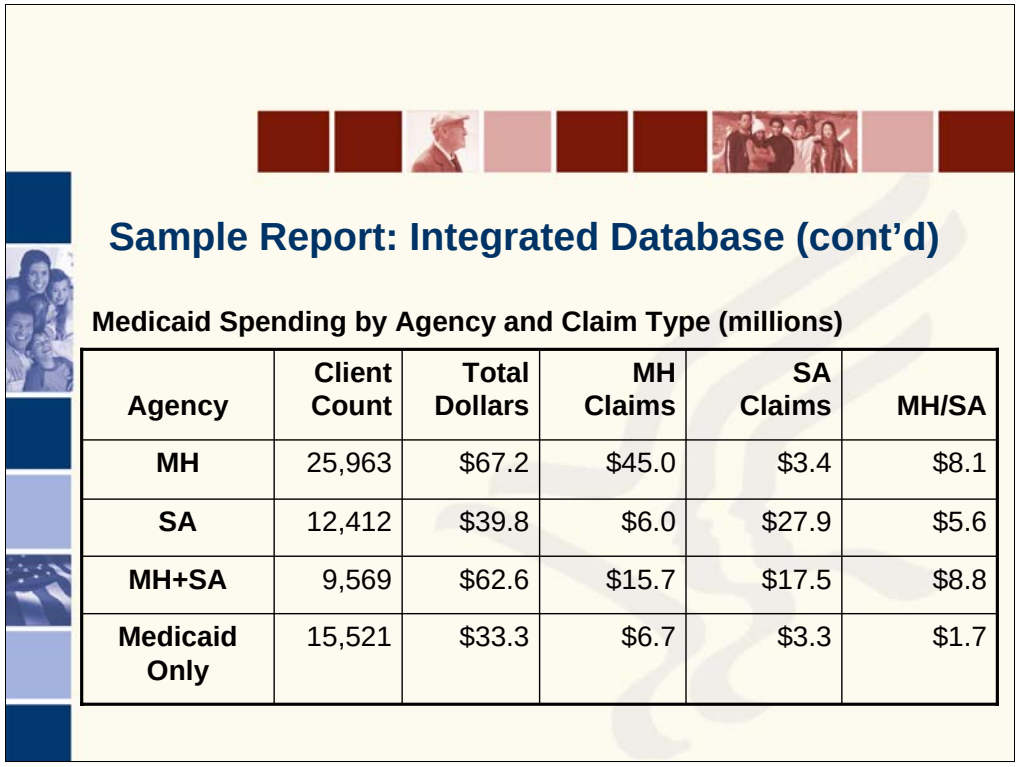


When the service-linked database is complete and tested, you have reached the final milestone. With truly integrated data, where you have linked clients *and* linked all available service records, there are numerous possibilities for reporting and analyses. Which subjects should be selected?

It makes sense to look back to the original compelling issues and questions that guided data integration. The issues that led to data integration will help you select the types of reporting and analysis. The database was built to address those questions.

Some issues are on-going, so you will probably want to establish a schedule of regular reporting – monthly, quarterly – that addresses on-going State or Agency issues. Other questions are better addressed with one-time analyses.

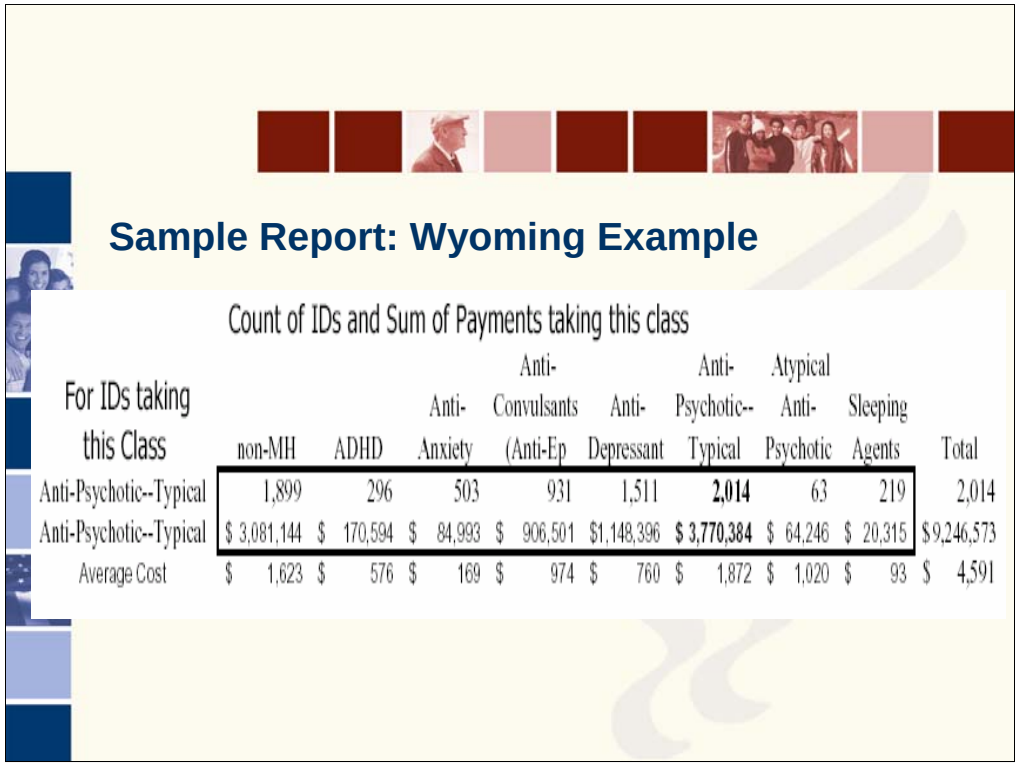
We suggest a couple of reports here to help stimulate ideas.



Here is an example of a good overview: Medicaid spending on clients with MH and/or SA conditions.

The report breaks down Medicaid costs by tracking agency – MH and SA are separate entities in this example. This is Medicaid spending, so by definition, everyone included in the report is serviced by Medicaid. The rows are mutually exclusive: clients appear in only one row. The first row represents Medicaid clients also in MH; the second row represents Medicaid clients also in SA; third row counts were tracked by Medicaid, MH, and SA; and the fourth row consists of Medicaid-only clients who have some MH or SA claim or condition.

After the header-column, the first column shows the number of clients in each category: MH and Medicaid, SA and Medicaid, and so on. The remaining columns reflect spending: Medicaid spending. This report does not account for any spending by MH or SA agencies. Total dollars are all Medicaid spending for the group; MH Claims are claims classified as MH because of diagnosis, procedure, or provider; SA Claims are similarly classified; and MH/SA consist of claims that are either both MH and SA, or cases in which it was not possible to distinguish between MH or SA.



Here is an example from Wyoming that examines Medicaid prescription drug spending for people with MH or SA conditions. This section – part of a larger report – shows clients taking typical anti-psychotics.

Here are some details to help you read the report:

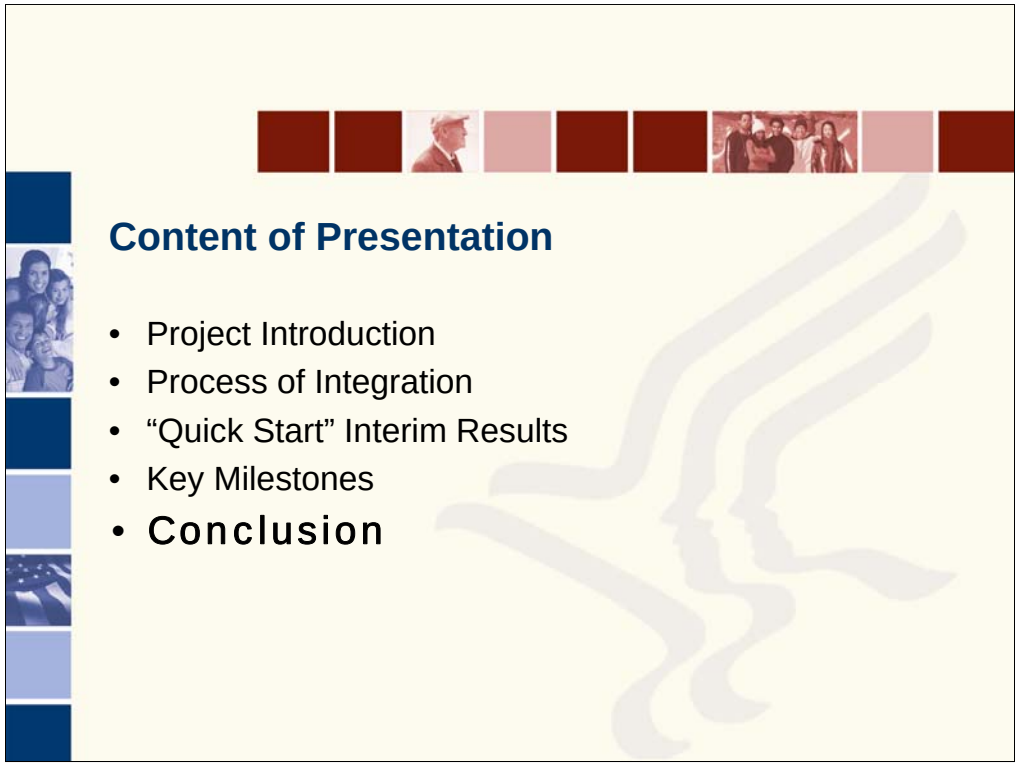
• The fourth column from the right – with the heading “typical anti-psychotics” – shows us that 2,014 people are taking typical anti-psychotics at a total cost of nearly $3.8 million, or $1,872 per person.

• The column on the far right reveals that total Medicaid prescription drug spending for this population of 2,014 was more than $9.2 million, or $4,591 per person.

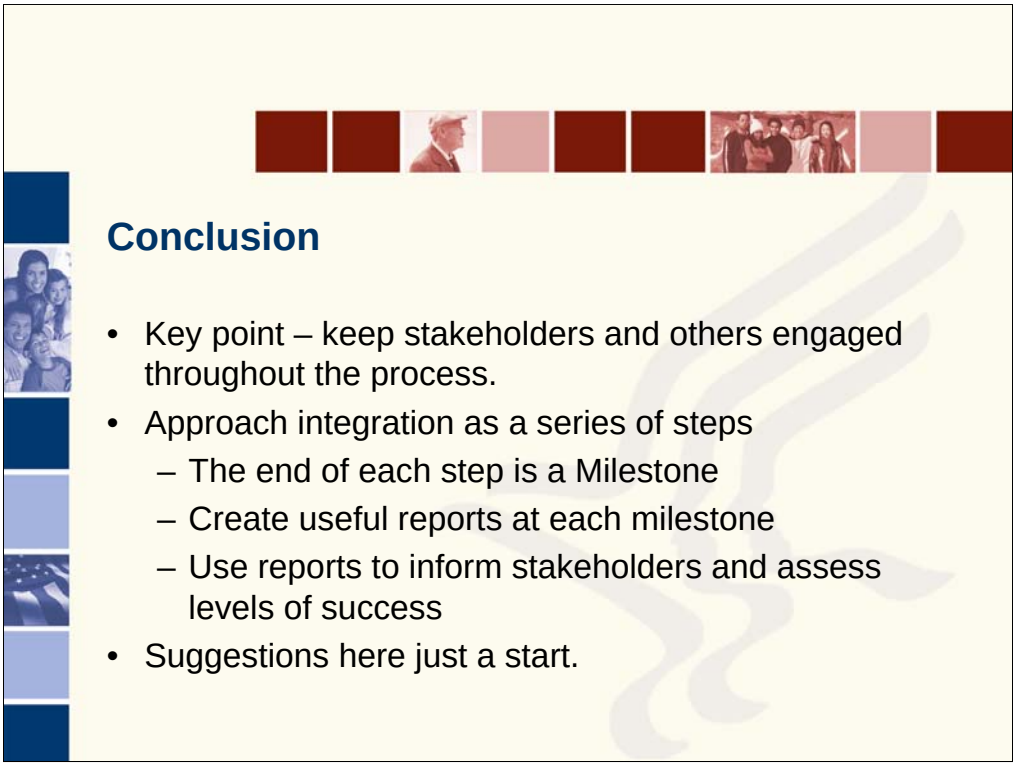
• The other columns indicate sub-groups that receive some other prescription, in addition to typical anti-psychotics. The columns indicate how many people are in the sub-group, the dollars spent on the particular drug class, and the average cost per person for that sub-group. To illustrate two examples:

On the left, the first column – with the heading “non-MH” – tells us that 1,899 of the 2,014 clients in this group also take some non-MH prescription, such as heart medication, at a cost of more than $3 million, or $1,623 per person.

The second column from the right – with the heading “Sleeping Agents” – tells us that 219 of the 2,014 clients have prescriptions for sleeping agents and the average cost is $93 per person for these 219 people.



We are approaching the end of our session.



We hope that this presentation has given you some useful ideas that will keep your data integration project moving. Remember that data integration can take a great deal of time, and it is important to engage stakeholders on a regular basis. Involving stakeholders throughout the integration process will facilitate stakeholder interest and support. This will encourage project momentum.

Start creating reports early in the process and keep generating information throughout the data integration effort. Use these reports to demonstrate progress and illuminate the importance of integrated data.

Tackle data integration as a series of steps, and create a visible product at the conclusion of each step. Tell a story with the data. Add descriptions and background to your reports so your audience understands what you are telling them. Use your results to push the integration process forward.

Finally, remember that these are only suggestions. Tailor your work to your specific situation and needs. Your actual reports will depend on issues unique to your state.

Thank you.